Statistical Inference

Worksheet 3:

1. The median age of the onset of diabetes is thought to be 45 years. The ages at onset of a random sample of 30 people with diabetes are:

35.5	44.5	39.8	33.3	51.4	51.3	30.5	48.9	42.1	40.3
46.8	38.0	40.1	36.8	39.3	65.4	42.6	42.8	59.8	52.4
26.2	60.9	45.6	27.1	47.3	36.6	55.6	45.1	52.2	43.5

   Assuming the distribution of the age of the onset of diabetes is symmetric, is there evidence to conclude that the median age of the onset of diabetes differs significantly from 45 years? Treat this as a paired two-sample test where the sample X is the given above and the pair to each point in X is 45. This variation of the Wilcoxon signed-rank test is a test for the median of the population.

2. [1][Exercise 16.1] A study of Darwin (1876), discussed in Hand et al. (1994) with the data in data(darwin), compared the ultimate heights of plants grown from otherwise comparable seedlings that were either cross-fertilized or self-fertilized. Compare the heights using the Wilcoxon signed-ranks procedure. If you are familiar with a t-test, would a paired t-test have been appropriate?

   To load the darwin data in `R` use the following lines:

```
install.packages("HH")
library(HH)
darwin
```

3. [1][Exercise 16.2] High levels of carbon monoxide transfer are a risk factor for contracting pneumonia. Ellis et al. (1987), also in Hand et al. (1994), studied the levels of carbon monoxide transfer in 7 chicken pox patients who were smokers. They were measured upon hospital admission and one week later. The data are in the file data(pox). This data set is also contained in the package "HH", follow the same lines as in the previous example.

   a. Verify the inappropriateness of a paired t-test for these data(Only if you are familiar with it). Discuss your reasoning.

   b. Analyze using the Wilcoxon signed-ranks procedure.

4. An obstetrician told a statistician, friend of his, that there were more births during the night (20.00-8.00) than during the day. The statistician contradicted this statement, saying that it just seemed so. To find out who was right, they recorded the time of occurrence of all spontaneous births under that doctor's care for a year. The results obtained were as follows:

Hours	2 – 5	5 – 8	8 – 11	11 – 14	14 – 17	17 – 20	20 – 23	23 – 2
in births	16	17	12	9	10	11	12	15

Do these results indicate that the statistician is right? Justify. (Assume that the significance level is kept at $\alpha = 0.05$).

5. Eight patients chosen at random underwent a certain treatment. Measurements of a parameter relevant to the study of the disease before the treatment, X, and after the treatment, Y, led to the following results:

sick	1	2	3	4	5	6	7	8
x (before)	62	42	74	37	51	43	80	82
y (after)	82	69	73	43	58	56	77	86

Does the treatment have an influence on the parameter studied? Justify.

6. An advertising agency claims that a particular product is at least 98% effective in removing shirt stains in 2 hours for any type of stain. In an experiment conducted by a consumer association, it was found that out of 100 shirts with stains, only 90 had no stains after two hours of product action. Should the advertising agency be sued, for a significance level of 2% ? Justify.

7. To analyse the breaking strength X of a certain type of yarn, an experiment was carried out with 10 samples of yarn, recording the following strengths (in grams): 295, 318, $305, 276, 297, 358, 342, 287, 345$ and 315.

   (a) Test the null hypothesis $\mathrm{H}_0 : \mathrm{Med(X)} = 300$ against the class of alternatives H_1 : $\mathrm{Med(X)} \neq 300$ for a level 0.05.

   (b) Manufacturing standards indicate that the resistance of a wire must be at least 310 g, with probability 0.9. In order to investigate whether the norms are respected, formulate and test the appropriate hypotheses, for a significance level of 0.05.

8. From the samples: .6, .8, 1.2, 1.4 and $1.2, 1.3, 1.3, 1.8, 2.4, 2.9$ from two distributions F and G, respectively, test the hypothesis $\mathrm{H}_0 : \mathrm{F(x)} \leq \mathrm{G(x)}$.

9. To compare two types of seed, A and B, a plot of land was divided into 9 lots with the same area and similar nature. In 4 lots chosen at random, seed A was used and in the others, seed B. The corresponding harvests (in “arrobas” 1 arroba = 15 Kg) were as follows:

X_A	42.0	44.5	46.0	47.0	
X_B	44.0	45.0	45.0	46.5	49.0

From these data and formulating the hypotheses that you deem necessary, compare seeds A and B.

References

[1] R. M. Heiberger and B. Holland. *Statistical analysis and data display.* Springer Texts in Statistics. Springer, New York, second edition, 2015. An intermediate course with examples in R.